Wielokategorialne systemy uczące się i ich zastosowanie w bioinformatyce

Rafał Grodzicki

Wielokategorialny system uczący się (multilabel learning system)

- Zbiór danych wejściowych:

  $$X = R^d$$

- Zbiór klas (kategorii):

  $$Y = \{1, 2, \ldots, Q\}$$

- Zbiór uczący:

  $$T = \{(x_1, Y_1), (x_2, Y_2), \ldots, (x_m, Y_m)\}, \qquad x_i \in X, \quad Y_i \subset Y \wedge Y_i \neq \emptyset$$

- Wielokategorialny system uczący się
  - Utworzenie wielokategorialnego klasyfikatora:

    $$h : X \rightarrow 2^Y$$

Wielokategorialny system uczący się (multilabel learning system)

- Zamiast klasyfikatora (h) system tworzy funkcję:

  $f : X \times Y \to R$

- Dla pary uczącej (x_i, Y_i), $x_i \in X$, $Y_i \subset Y \wedge Y_i \neq \emptyset$

  system dąży do generowania funkcji spełniającej

  warunek $(\forall y_1 \in Y_i \wedge y_2 \notin Y_i)$ $f(x_i, y_1) > f(x_i, y_2)$

- Na podstawie utworzonej funkcji można wygenerować klasyfikator:

  $(\forall x \in X)$ $h(x) = \{y \in Y : f(x, y) > t(x)\}$

  gdzie $t : X \to R$ jest funkcją progową

Problem wielokategorialnej klasyfikacji

1. Dekompozycja na niezależne problemy klasyfikacji binarnej
   - Nie uwzględnia korelacji pomiędzy różnymi klasami
2. Każdy podzbiór kategorii stanowi oddzielną klasę
   - Generuje dużą liczbę klas

     (Q kategorii – 2^Q klas)

Problem wielokategorialnej klasyfikacji

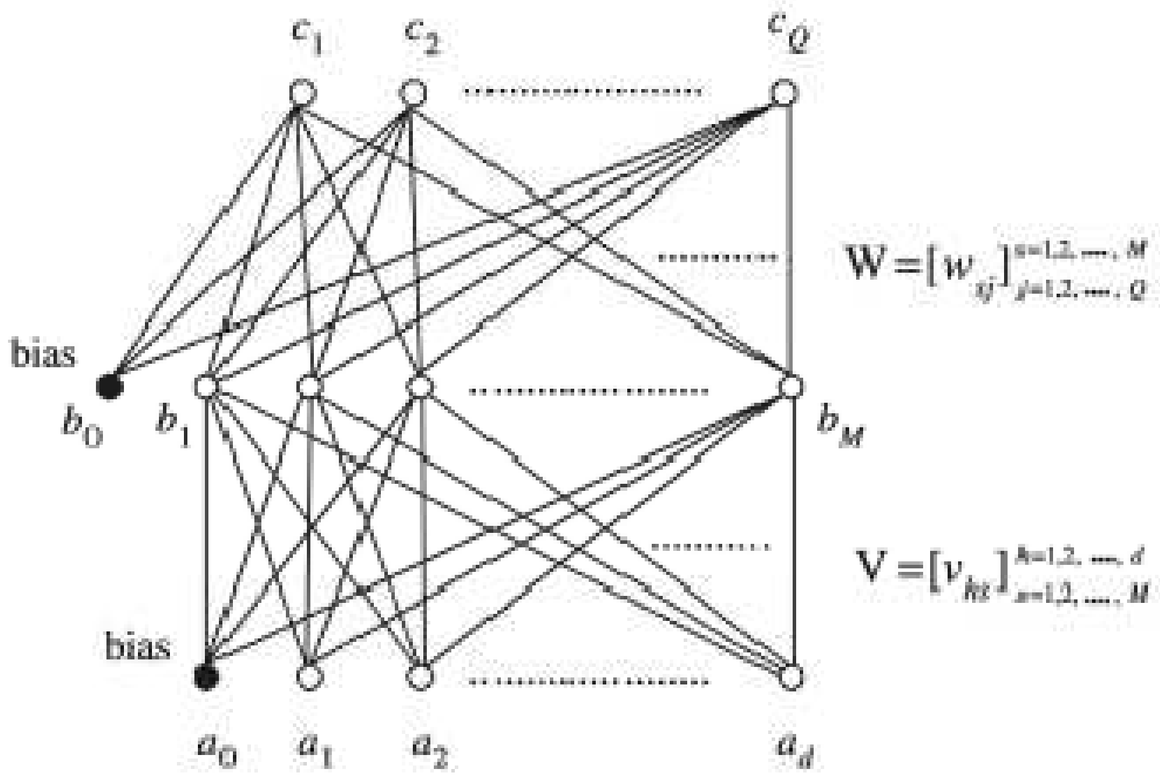
- Dobry wielokategorialny system uczący się
  - Uwzględnianie korelacji pomiędzy różnymi klasami (kategoriami)
  - Zachowanie małej liczby klas

Sieć neuronowa jako wielokategorialny klasyfikator

- BP-MLL (Backpropagation for Multilabel Learning)
- Autorzy:
  - Min-Ling Zhang
  - Zhi-Hua Zhou
- Pierwszy wielokategorialny system uczący się oparty na sieciach neuronowych
- Perceptron ze zmodyfikowaną funkcją błędu
- Uczenie – wsteczna propagacja błędu

Sieć neuronowa
jako wielokategorialny klasyfikator

- Architektura



Sieć neuronowa jako wielokategorialny klasyfikator

- Funkcja błędu
  - Klasyczna – błąd średniokwadratowy

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \sum_{j=1}^{Q} \left(c_j^i - d_j^i\right)^2, \quad c_j^i = c_j(x_i), \quad d_j^i = \begin{cases} +1, & j \in Y_i \\ -1, & j \notin Y_i \end{cases}$$

    - Uwzględnia poszczególne kategorie niezależnie
    - Nie uwzględnia korelacji pomiędzy kategoriami (klasami)
  - Na wyjściu sieci powinny być większe wartości dla kategorii należących do Y_i niż dla kategorii spoza Y_i

Sieć neuronowa jako wielokategorialny klasyfikator

- Funkcja błędu
  - Zmodyfikowana

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \frac{1}{|Y_i||\overline{Y}_i|} \sum_{(k,l)\in Y_i \times \overline{Y}_i} e^{-\left(c_k^i - c_l^i\right)}, \quad c_j^i = c_j(x_i)$$

- Koncentracja na różnicy pomiędzy wartością wyjściowymi dla kategorii należących do Y_i a wartościami wyjściowy dla kategorii spoza Y_i
- Silne karanie w przypadku wartości wyjściowych dla kategorii spoza Y_i większych niż dla kategorii z Y_i
- Uwzględnia zależności pomiędzy różnymi klasami – większe wartości na wyjściu sieci dla kategorii należących do Y_i niż dla kategorii spoza Y_i

Sieć neuronowa jako wielokategorialny klasyfikator

- Funkcja błędu
  - Przykład
    - Oczekiwane wartości na wyjściu sieci: 1, 1, -1

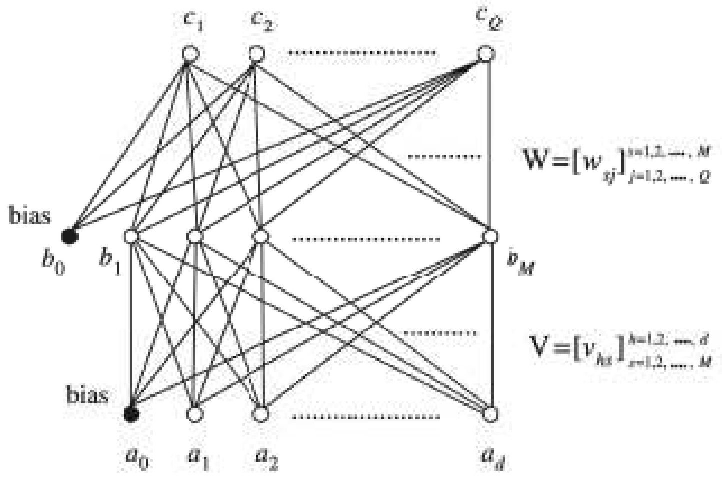
Rzeczywiste wartości na wyjściu sieci	**Klasyczna funkcja błędu**	**Zmodyfikowana funkcja błędu**
0.05, 0.05, -0.05	2.7075	0.9048
0.3, 0.3, 0.3	2.67	1

Sieć neuronowa jako wielokategorialny klasyfikator

- Uczenie
  - Algorytm wstecznej propagacji błędu
  - Modyfikacje wag (warstwy ukryta – wyjściowa):

$$\Delta w_{sj} = -\mu \frac{\partial E_i}{\partial w_{sj}} = \mu d_j b_s$$

$$d_j = \begin{cases} \left(\frac{1}{|Y_i||\overline{Y}_i|} \sum_{l \in \overline{Y}_i} e^{-(c_j - c_l)} \right)(1 + c_j)(1 - c_j), & j \in Y_i \\ \left(-\frac{1}{|Y_i||\overline{Y}_i|} \sum_{k \in Y_i} e^{-(c_k - c_j)} \right)(1 + c_j)(1 - c_j), & j \in \overline{Y}_i \end{cases}$$

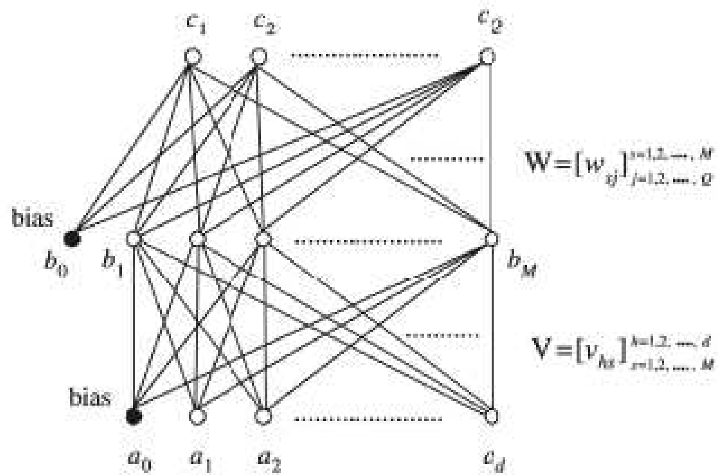


Sieć neuronowa
jako wielokategorialny klasyfikator

- Uczenie
  - Algorytm wstecznej propagacji błędu
  - Modyfikacje wag (warstwy wejściowa – ukryta):

$$\Delta v_{hs} = -\mu \frac{\partial E_i}{\partial v_{hs}} = \mu e_s a_h$$

$$e_s = \left(\sum_{j=1}^{Q} d_j w_{sj} \right) (1 + b_s)(1 - b_s)$$



Sieć neuronowa jako wielokategorialny klasyfikator

- Klasyfikacja
  - Na podstawie wartości wyjściowych sieci ustalany jest zbiór kategorii (klas) odpowiadający danym wejściowym:

    $\{j \in Y : c_j > t(x)\}$

    gdzie $t : X \rightarrow R$ jest funkcją progową
  - Funkcja progowa:
    - Stała funkcja (t(x) = 0)
    - Wyznaczana na podstawie zbioru uczącego

Sieć neuronowa jako wielokategorialny klasyfikator

- Klasyfikacja
  - Funkcja progowa – wyznaczanie na podstawie zbioru uczącego:
    - Definicja: $t(x) = w^T c(x) + b, \quad c(x)^T = (c_1(x), c_2(x), \ldots, c_Q(x))$
    - Dla każdej pary uczącej $(x_i, Y_i), \quad x_i \in X, \quad Y_i \subset Y \wedge Y_i \neq \emptyset$ określona jest wartość funkcji progowej:
      $t(x_i) = \arg\min_t \left(\left|\left\{ k \in Y_i : c_k^i \leq t \right\}\right| + \left|\left\{ l \in \overline{Y_i} : c_l^i \geq t \right\}\right| \right)$
    - Parametry funkcji progowej wyznaczane na podstawie rozwiązania równania:
      $Aw' = t,$
      $A[i] = (c_1^i, c_2^i, \ldots, c_Q^i, 1), \quad w' = (w^T, b)^T, \quad t = (t(x_1), t(x_2), \ldots, t(x_m))$

Miary oceny jakości klasyfikacji wielokategorialnej

- Hamming loss
  - Określa jak często występuje błędna klasyfikacja
  - Im mniejsza wartość tym lepiej

$$hloss_S(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{Q}\left|h(x_i)\Delta Y_i\right|,$$

$$h(x_i)\Delta Y_i = \left(h(x_i)\cup Y_i\right)\setminus\left(h(x_i)\cap Y_i\right)$$

Miary oceny jakości klasyfikacji wielokategorialnej

- One-error
  - Określa jak często kategoria o najwyższej wartości wyjściowej nie należy do zbioru Y_i
  - Im mniejsza wartość tym lepiej

$$oneerror_S(f) = \frac{1}{p} \sum_{i=1}^{p} \left(\left[\left(\arg\max_{y \in Y} f(x_i, y) \right) \notin Y_i \right] ? 1 : 0 \right)$$

Miary oceny jakości klasyfikacji wielokategorialnej

- Ranking loss
  - Określa uśrednioną część par kategorii
    $$(y_1, y_2), \quad (y_1 \in Y_i \wedge y_2 \notin Y_i \wedge f(x_i, y_1) \leq f(x_i, y_2))$$
  - Im mniejsza wartość tym lepiej

$$rloss_S(f) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{|Y_i||\overline{Y_i}|} \left| \left\{ (y_1, y_2) \in Y_i \times \overline{Y_i} : f(x_i, y_1) \leq f(x_i, y_2) \right\} \right|$$

Klasyfikacja wielokategorialna – zastosowanie w bioinformatyce

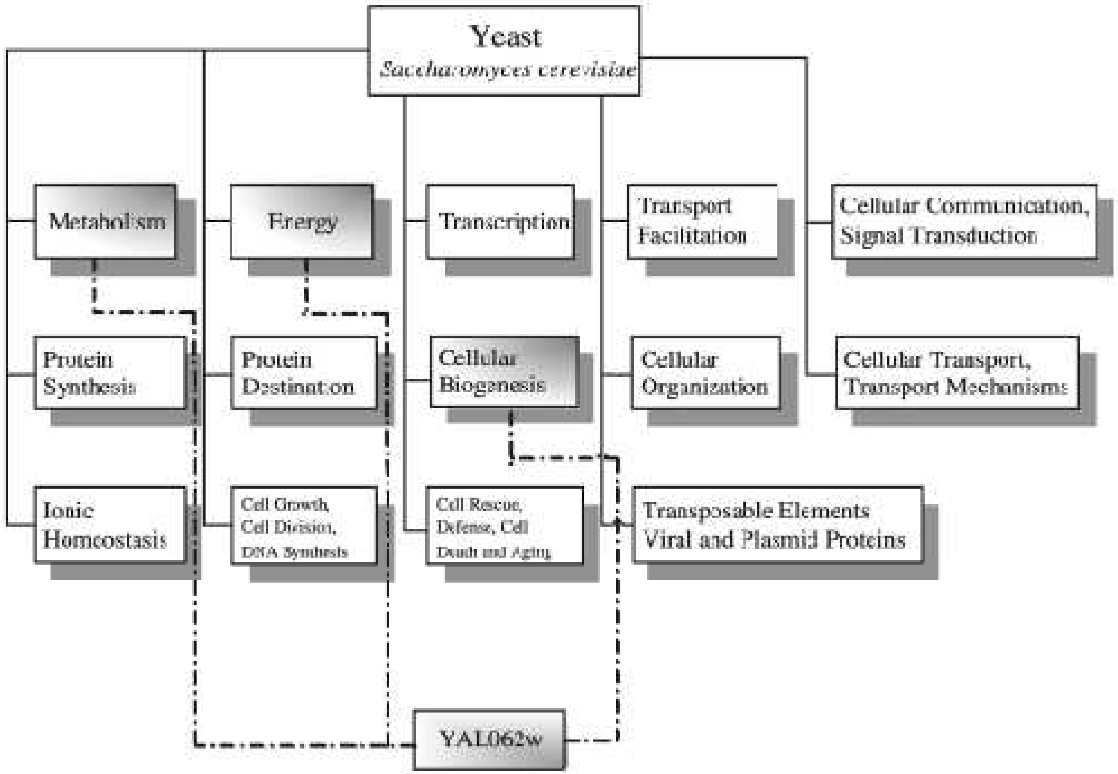
- Genomika funkcjonalna (functional genomics)
  - Cele:
    - Określenie funkcji genów i kodowanych przez nie białek
    - Poznanie procesów zachodzących w organizmach żywych
  - Narzędzia:
    - Mikromacierze DNA
      - Poziomy ekspresji genów w różnych warunkach
    - Sekwencje
      - nukleotydów w danym genie
      - aminokwasów w białku kodowanym przez gen
    - Profile filogenetyczne
      - Ciąg bitów odpowiadających genomom różnych gatunków
      - 1 – gen występuje w danym genomie, 0 – w p.p.

Klasyfikacja wielokategorialna – zastosowanie w bioinformatyce

- Genomika funkcjonalna (functional genomics)
  - Problem klasyfikacji wielokategorialnej:
    - Każdy gen powiązany ze zbiorem funkcji (klas)
    - Przykład – genom drożdży:
      - Określone 14 klas funkcji genów
      - Gen YAL062w należy do klas:
        - Metabolism
        - Energy
        - Cellular Biogenesis

Klasyfikacja wielokategorialna – zastosowanie w bioinformatyce

- Klasyfikacja wielokategorialna w genomie drożdży



Wyznaczanie klas fukcjonalnych genomu drożdży (za pomocą BP-MLL)

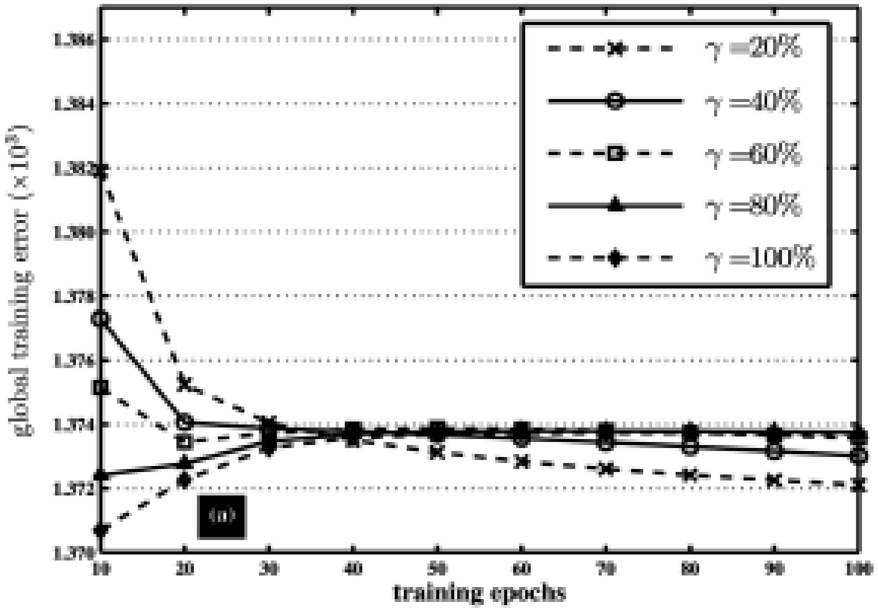
- Dane wejściowe
  - Profile ekspresji genów (z mikromacierzy DNA)
  - Profile filogenetyczne
  - 103 wymiarowy wektor
- Zbiór klas funkcjonalnych
  - Struktura hierarchiczna – 4 poziomy
  - 1. poziom – 14 klas funkcjonalnych
  - Każdy gen powiązany z wieloma klasami:
    - średnia: 4.24
    - odchylenie standardowe: 1.57
- Zbiór uczący
  - 2417 genów powiązanych z klasami funkcjonalnymi

Wyznaczanie klas fukcjonalnych genomu drożdży (za pomocą BP-MLL)

- Parametry sieci neuronowej
  - Współczynnik uczenia: 0.05
  - Liczba neuronów w warstwie ukrytej:
    - Od 20 do 100 procent liczby elementów wejściowych (krok: 20 procent)
  - 100 epok
- Przebieg uczenia
  - Walidacja krzyżowa
    - Zbiór danych losowo dzielony na 10 równych części
    - 10 cykli uczenia
      - 1 cześć – zbiór testowy
      - 9 części – zbiór uczący
      - Wyznaczenie miar oceny jakości klasyfikacji
    - Wyznaczenie uśrednionych (z 10 cykli) miar oceny jakości klasyfikacji

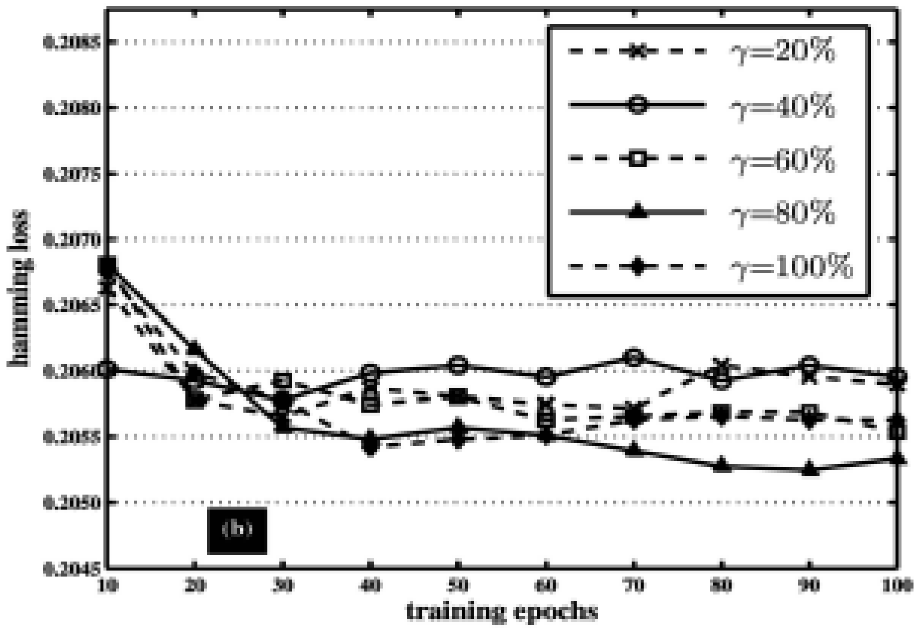
Wyznaczanie klas fukcjonalnych genomu drożdży (za pomocą BP-MLL)

- Miary jakości:
  - Błąd globalny



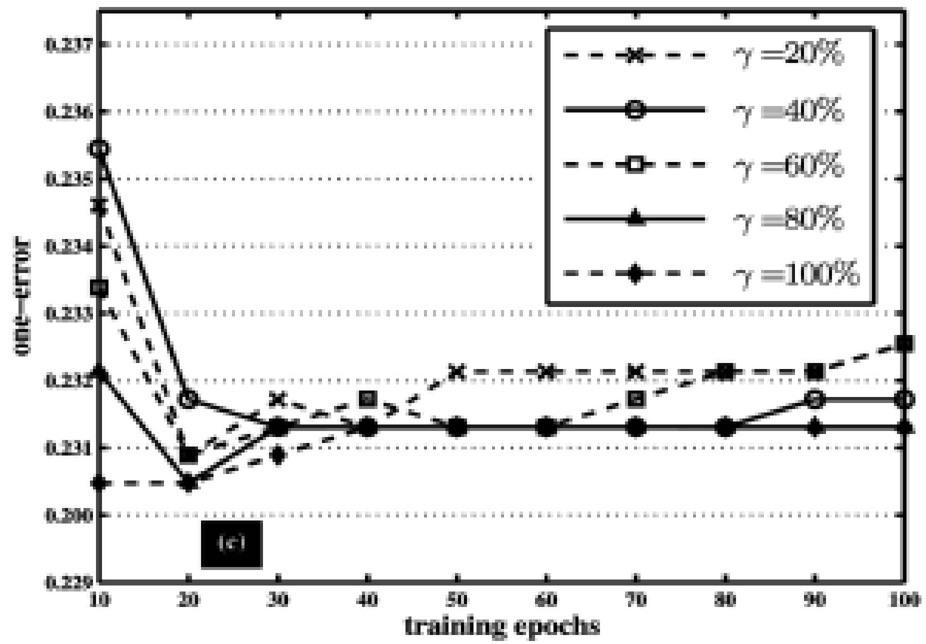
Wyznaczanie klas fukcjonalnych genomu drożdży (za pomocą BP-MLL)

- Miary jakości:
  - Hamming loss



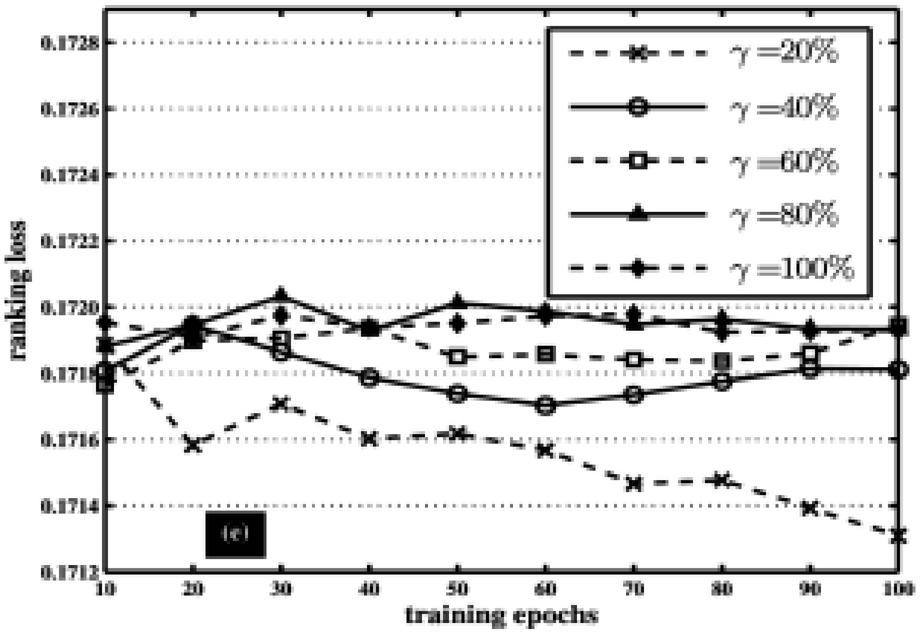
Wyznaczanie klas fukcjonalnych genomu drożdży (za pomocą BP-MLL)

- Miary jakości:
  - One-error



Wyznaczanie klas fukcjonalnych genomu drożdży (za pomocą BP-MLL)

- Miary jakości:
  - Ranking loss



Wyznaczanie klas fukcjonalnych genomu drożdży (za pomocą BP-MLL)

- BP-MLL a klasyczny perceptron

Miara jakości klasyfikacji	**BP-MLL**	**Klasyczny perceptron**
Hamming loss	0.206 (±0.011)	0.209 (±0.008)
One-error	0.233 (±0.034)	0.245 (±0.032)
Ranking loss	0.171 (±0.015)	0.184 (±0.017)

  – Liczba neuronów w warstwie ukrytej – 20% liczby elementów wejściowych

Propozycje eksperymentów badawczych

- Zastąpienie perceptronu siecią radialną
  - Warstwa ukryta – neurony radialne opisane np. funkcją Gaussa:

$$\varphi\left(x; c, \sigma^2\right) = e^{-\left(\frac{\|x-c\|}{2\sigma^2}\right)}$$

  - Uczenie
    - Parametry neuronów warstwy ukrytej – radialnych
      - uczenie bez nadzoru – klasteryzacja (np. k-means)
    - Wagi połączeń (warstwy ukryta – wyjściowa)
      - Wsteczna propagacja błędu

Propozycje eksperymentów badawczych

- Modyfikacje funkcji błędu

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \max_{(k,l) \in Y_i \times \overline{Y}} \left(e^{-\left(c_k^i - c_l^i\right)} \right), \quad c_j^i = c_j\left(x_i\right)$$

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \frac{1}{\left|\min\left(Y_i, K\right)\right| \left|\max\left(\overline{Y}_i, K\right)\right|} \sum_{(k,l) \in \min\left(Y_i, K\right) \times \max\left(\overline{Y}_i, K\right)} e^{-\left(c_k^i - c_l^i\right)}, \quad c_j^i = c_j\left(x_i\right)$$

Propozycje eksperymentów badawczych

- Modyfikacje funkcji progowej
  - Uogólnienie
  - Zależność od:
    - wektora wejściowego
    - kategorii (klas)
  - Adaptacja parametrów funkcji progowej na podstawie zbioru uczącego

Bibliografia

- Min-Ling Zhang, Zhi-Hua Zhou - "Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 10, 2006
- A. Clare, "Machine Learning and Data Mining for Yeast Functional Genomics", PhD dissertation, Dept. of Computer Science, Univ. of Wales Aberystwyth, 2003
- A. Elisseeff, J. Weston - "A Kernel Method for Multi-Labelled Classification", Advances in Neural Information Processing Systems, vol. 14, pp. 681-687, 2002
- A. Clare, R.D. King, "Knowledge Discovery in Multi-Label Phenotype Data", Lecture Notes in Computer Science, vol. 2168, pp. 42-53, Berlin: Springer, 2001
- P. Pavlidis, J. Weston, J. Cai, and W.N. Grundy, "Combining Microarray Expression Data and Phylogenetic Profiles to Learn Functional Categories Using Support Vector Machines", Proc. Fifth Ann. Int’l Conf. Computational Molecular Biology (RECOMB ’01), pp. 242-248, 2001